GP aka Kriging (MLclass SUPAERO 2018)

Prof. Joseph Morlier

1/ First Hour (GP)

2/ Second Hour (SBO)

1/ First Hour (GP)

2/ Second Hour (SBO)

A bit of History

Kriging (Pionneer)	Gaussian Processes (link with AI)
Developed by Daniel Krige – 1951; formalized by Georges Mathéron in the 60's (Mines Paris)	Neural network with infinite neurons tend to Gaussian Process 1994
Evaluation: minimize error variance	Evaluation: Marginal Likelihood

Krige, D. G., 1951, A statistical approach to some basic mine valuation problems on the Witwatersrand: J. Chem. Metal. Min. Soc. South Africa, v. 52, p. 119-139.

Matheron, G., 1963b, Principles of geostatistics: Economic Geol., v. 58, p. 1246-1266.

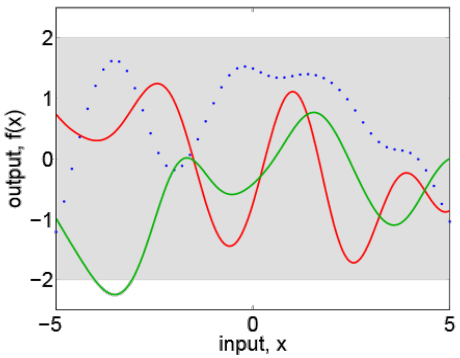
Neal, R. Priors for infinite networks. Tech. rep., University of Toronto, 1994.

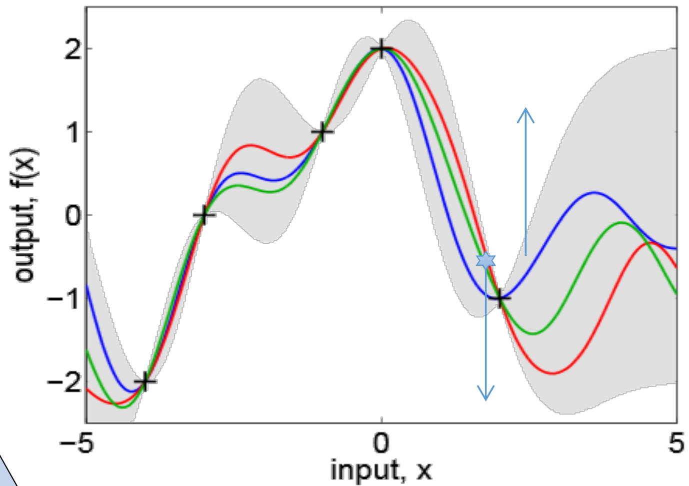
Williams, C. K. I., and Rasmussen, C. E. Gaussian processes for regression. *Advances in Neural Information Processing Systems 8* (1996), 514–520.

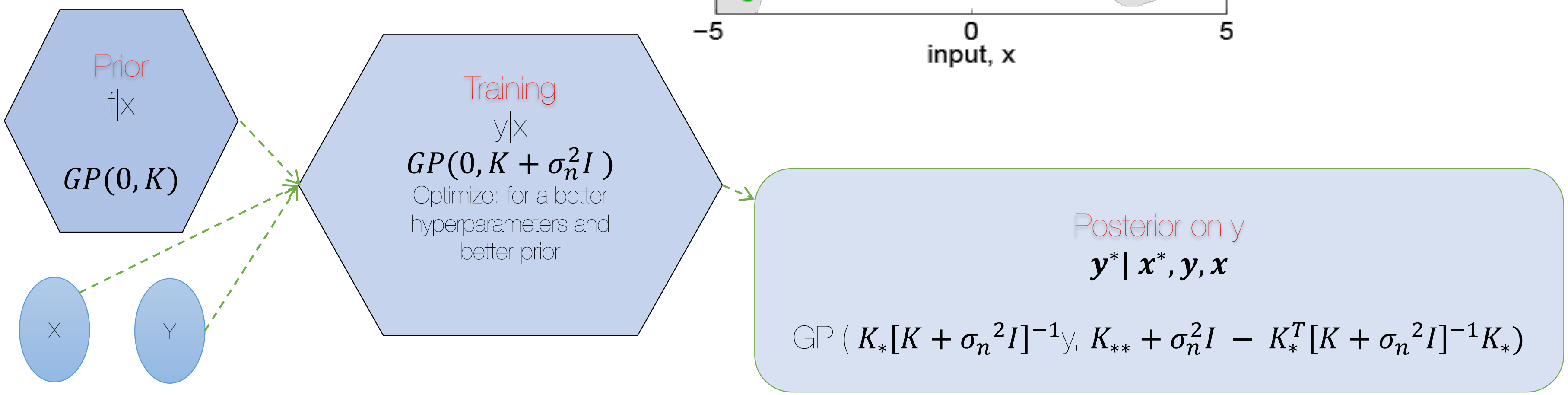
http://extrapolated-art.com

Gaussian Process Regression

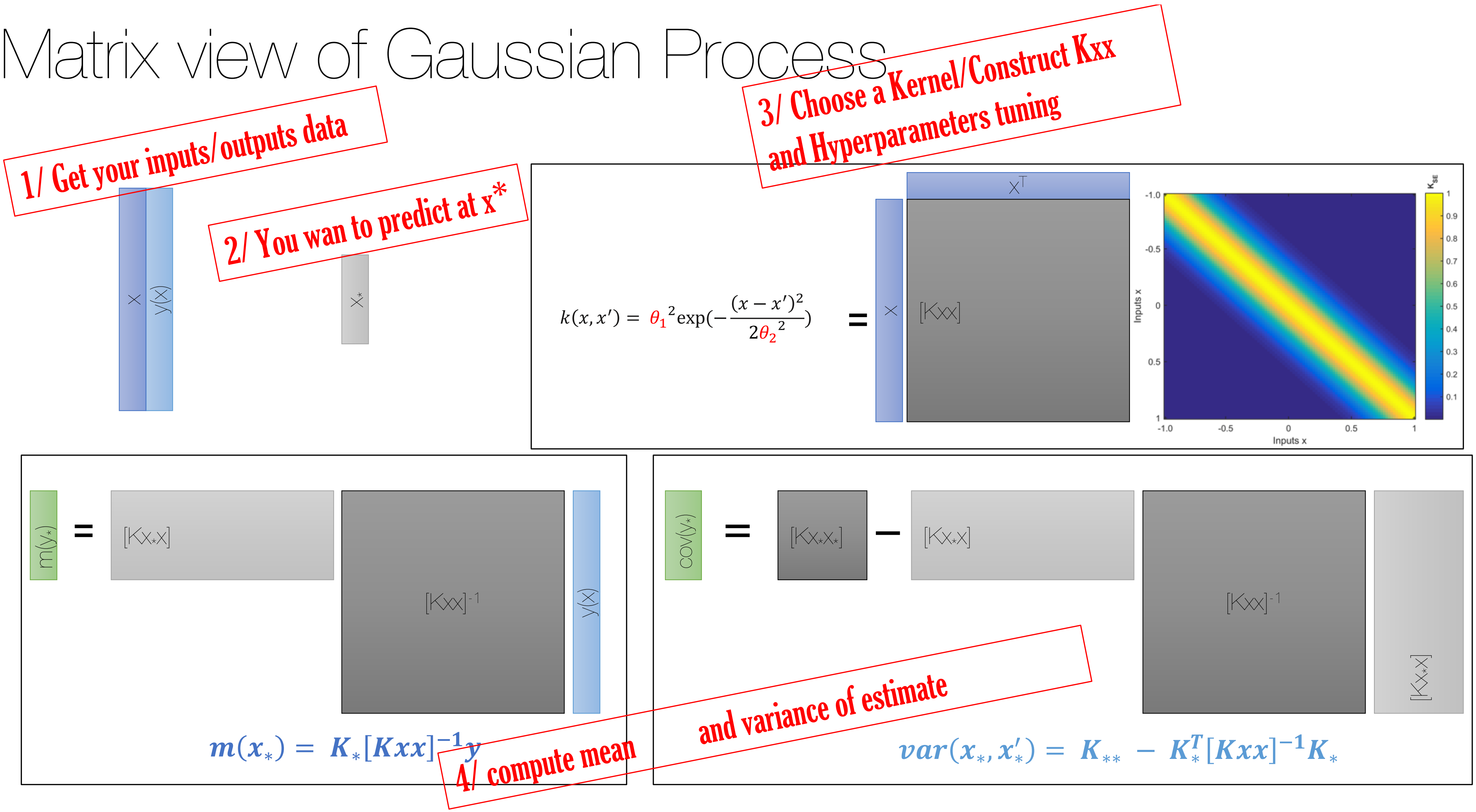
Image Source: http://mlg.eng.cam.ac.uk/teaching/4f13/1314/







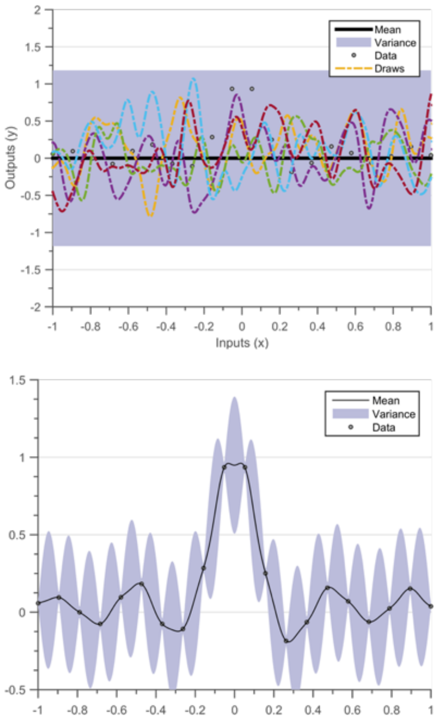
Matrix view of Gaussian Process



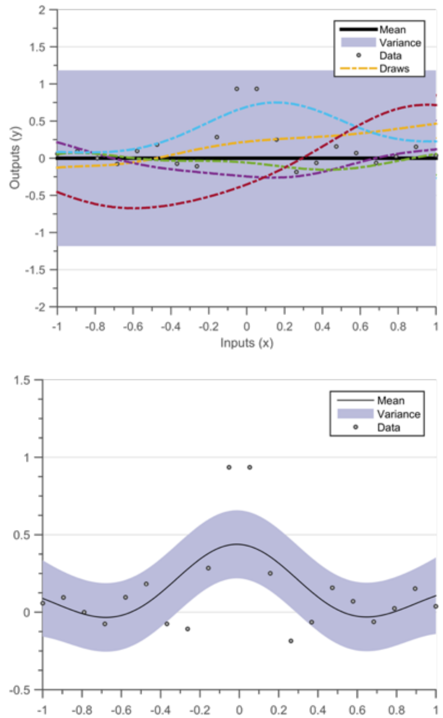
Optimizing Marginal Likelihood (ML)

$$\text{ML} = log(p(y|X,\theta)) = -\frac{1}{2}y^T K^{-1} y - \frac{1}{2} log|K| - \frac{n}{2} log(2\pi)$$

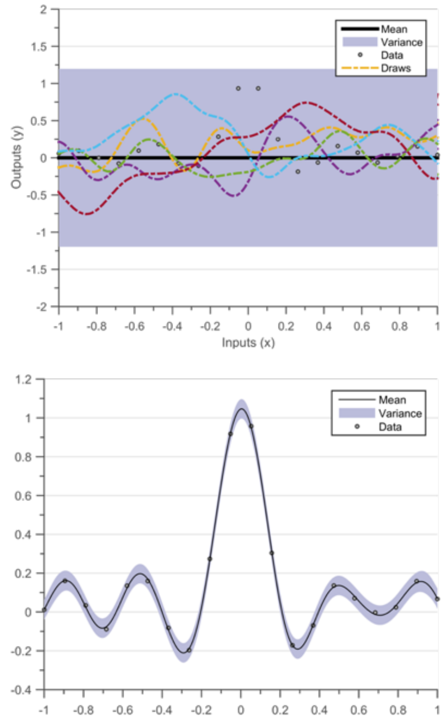
- It is a combination of **data-fit term**, a **complexity penalty** term and a **normalization term**



ML = -8.2



ML = -35.3



ML = 6,04

Hyperparameters tuning

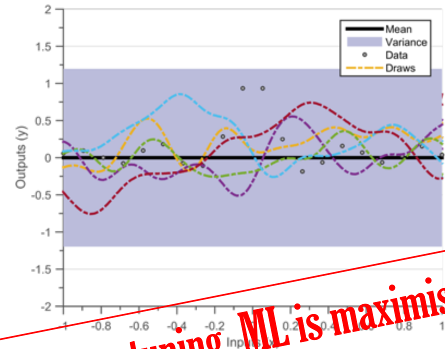
$$k(x, x') = \theta_1{}^2 \exp(-\frac{(x - x')^2}{2\theta_2{}^2})$$

Only two hyperparameters:

→ The lengthscale θ_2 or ℓ determines the length of the 'wiggles' in your function.

→ The output variance $\theta_1{}^2$ or σ^2 determines the average distance of your function away from its mean. It's just a scale factor.

→ A third hyperparameter θ_3 or σ_n^2 is often used (noise) $GP(0, K + \sigma_n^2 I)$



3/ Hyperparameters tuning. ML is maximised, θ^* is found

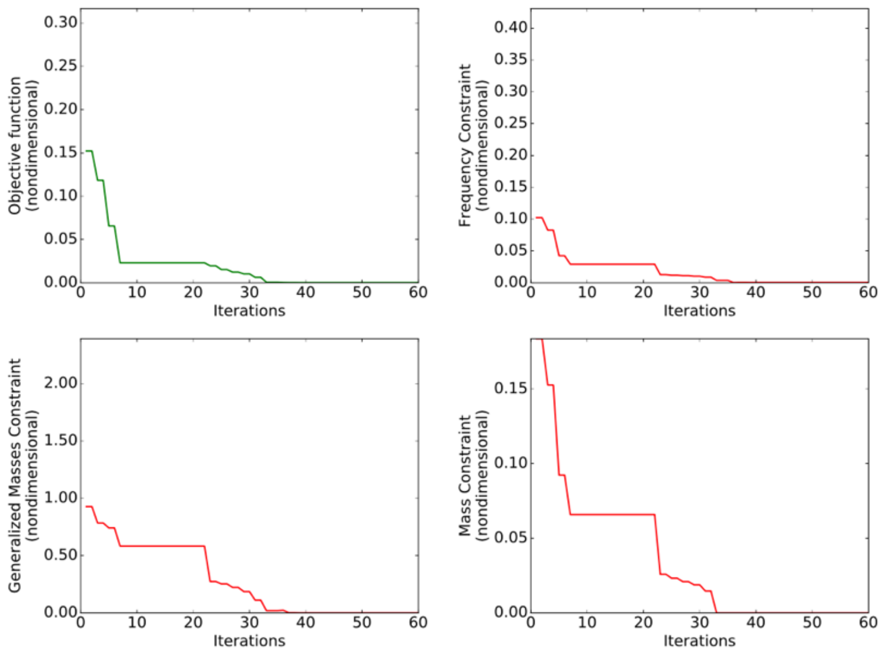
1/ First Hour (GP)

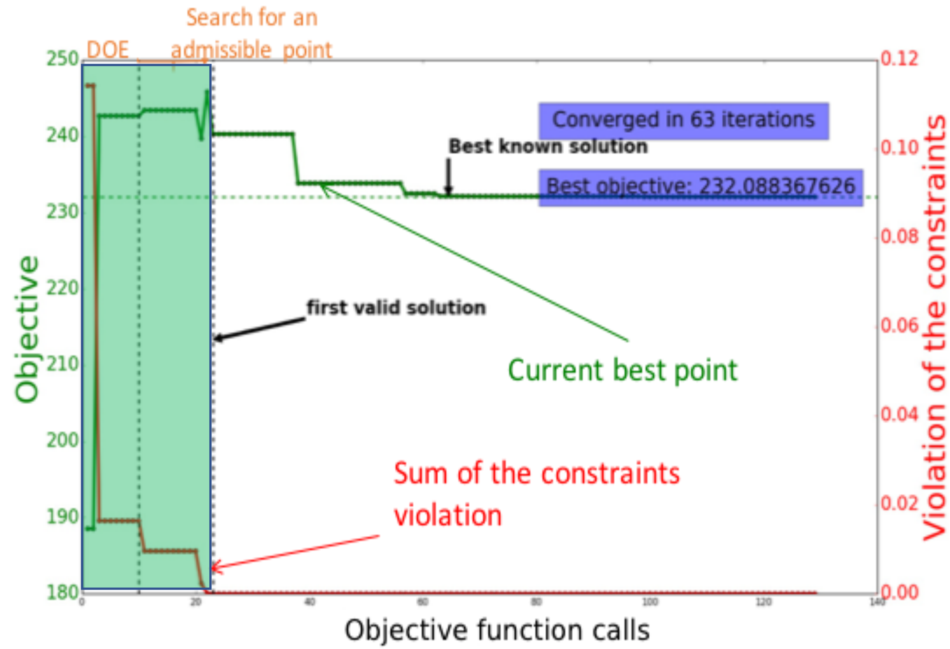
2/ Second Hour (SBO)

New paradigm for Surrogate Based Optimization (SBO)

Gradient based Optimality, Feasibility

SBO Exploration, Exploitation





Stopping criteria: tolfun, tolx, maxiter

Stopping criteria: Max Budget (Function calls)

The goal is: find min of f(x) by sampling + and Kriging updating

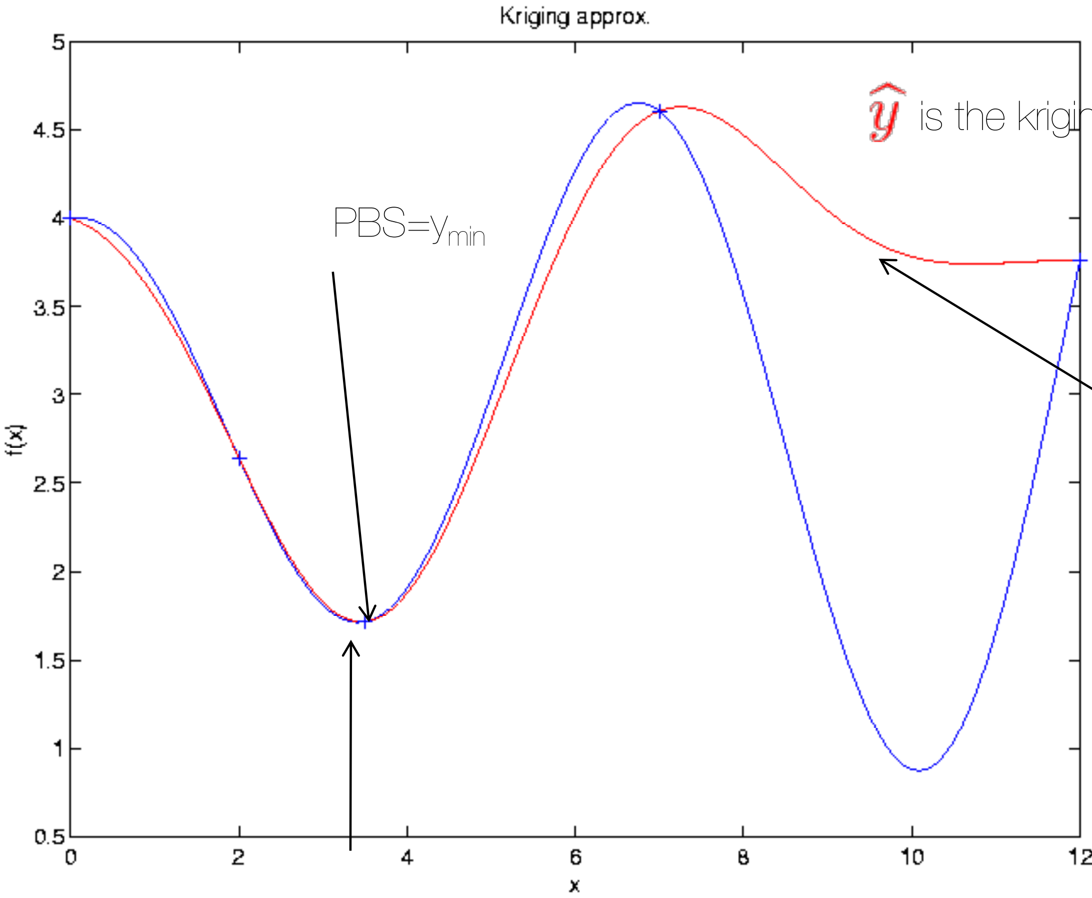
Where do I need to update my sampling?

We note the present best solution ($PBS=y_{min}$)

At every x there is some chance of improving on the PBS.

Then we ask: Assuming an improvement over the PBS, where is it likely be largest?

$\hat{y}$ is the kriging prediction of y



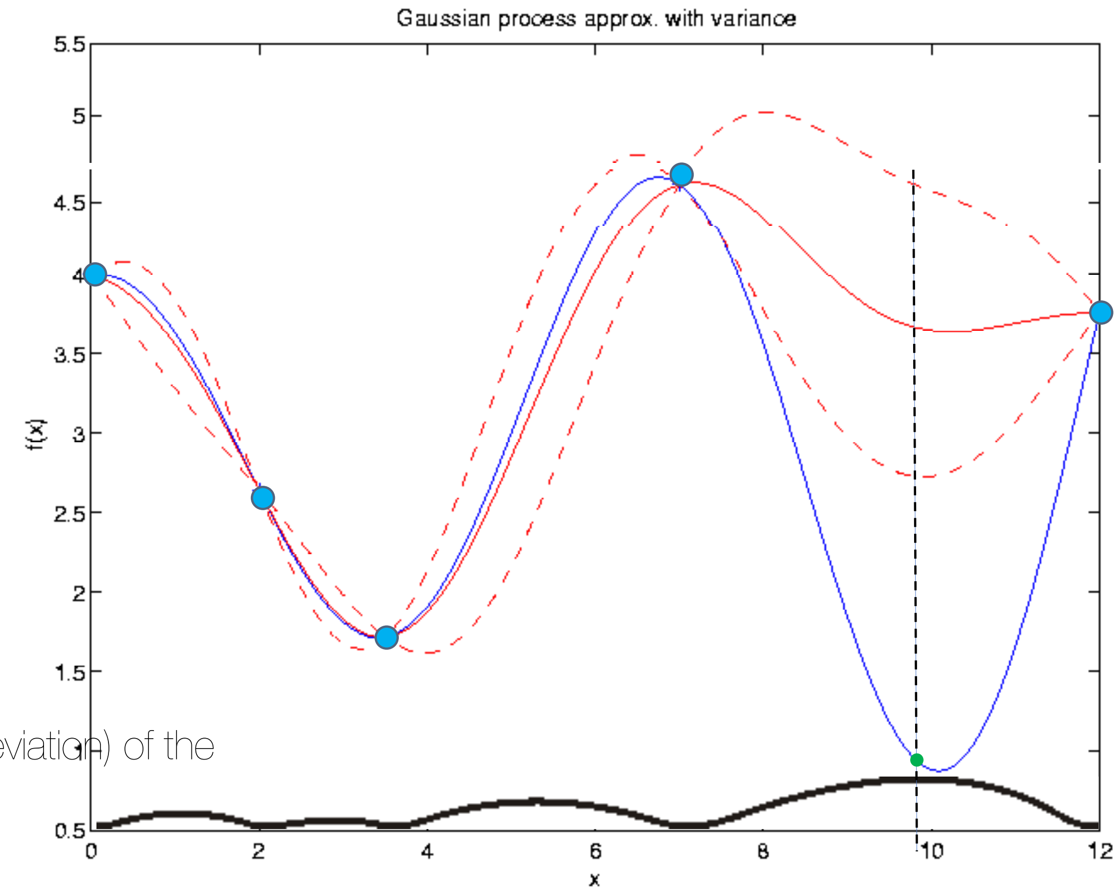
Need points here ?
Exploration

Exploitation may drive the optimization to a local optimum

In supervised mode … have a look to max(RMSE)

Not possible to compute the error: we don't know for each x the true value of the function

But…. Can we use GP properties ?

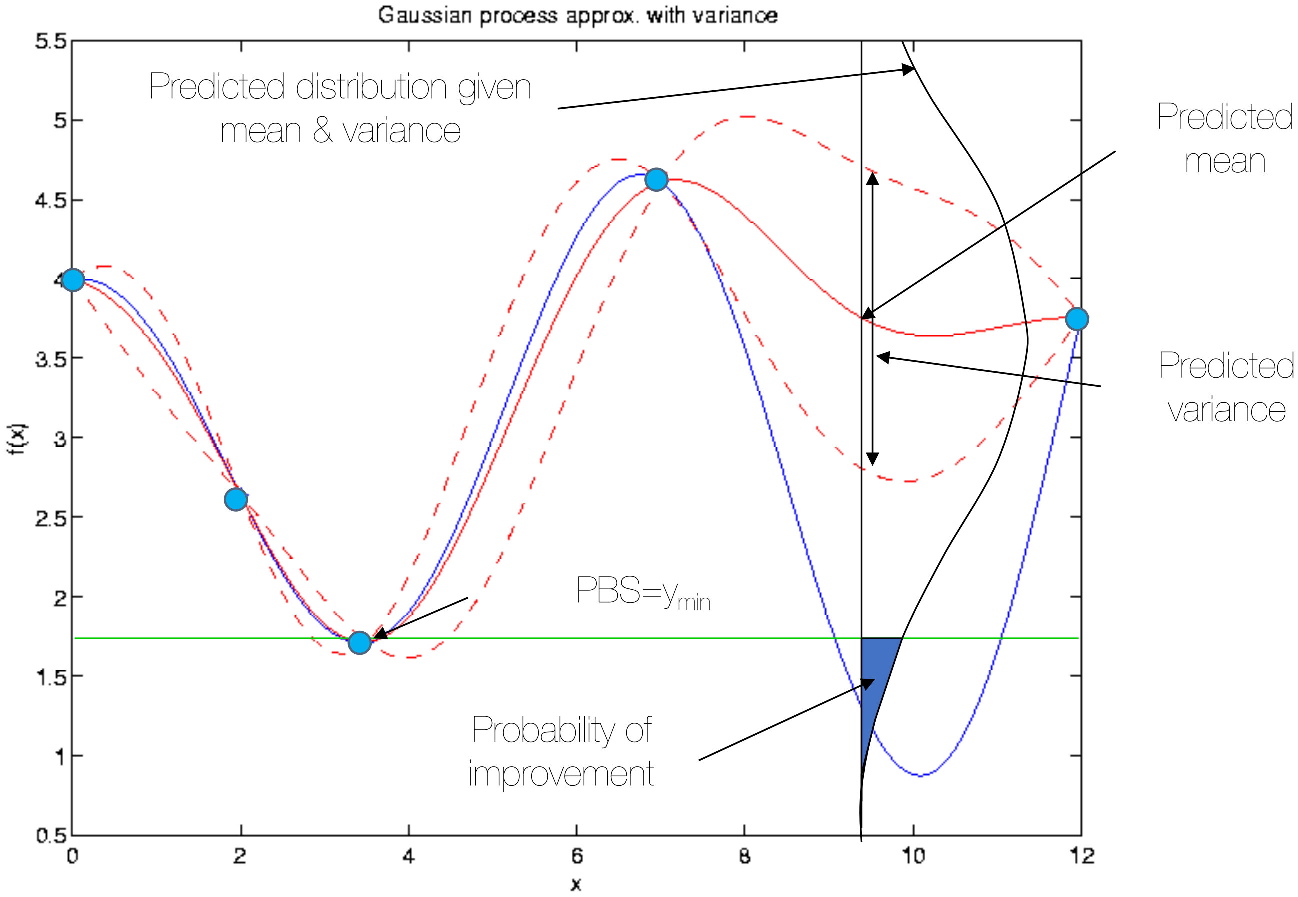


$\hat{y}$ is the kriging prediction of y

$\hat{s}$ is the estimation error (standard deviation) of the prediction (often noted σ_y)

PBS=y_{min}

Probability of improvement



Improvement . . . explicitely

- *Improvement* : $I(\mathbf{x}) = \max\left(y_{min} - \hat{Y}(\mathbf{x}), 0\right)$
- *Expected Improvement* :

$$\mathrm{EI}(x) = \mathrm{E}[\max(0, y_{\min} - \hat{y}(x))]$$

$$E[I(\mathbf{x})] = \int_{-\infty}^{y_{min}} (y_{min} - \hat{y})\varphi\left(\frac{y_{min} - \mu_{\hat{Y}}(\mathbf{x})}{\sigma_{\hat{Y}}(\mathbf{x})}\right) d\hat{y}$$

$$E[I(\mathbf{x})] = (y_{min} - \mu_{\hat{Y}}(\mathbf{x}))\Phi\left(\frac{y_{min} - \mu_{\hat{Y}}(\mathbf{x})}{\sigma_{\hat{Y}}(\mathbf{x})}\right) + \sigma_{\hat{Y}}(\mathbf{x})\varphi\left(\frac{y_{min} - \mu_{\hat{Y}}(\mathbf{x})}{\sigma_{\hat{Y}}(\mathbf{x})}\right)$$

Exploitation

Exploration

global optimum can be found because P[I(x)] = 0 when s = 0 so that there is no probability of improvement at a point which has already been sampled → guarantees global convergence

Φ: cumulative distribution function $\mathcal{N}(0,1)$ ϕ: probability density function $\mathcal{N}(0,1)$

***Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. Journal of Global optimization, 13(4), 455-492.**

Infill Criteria : max(Expected improvement)

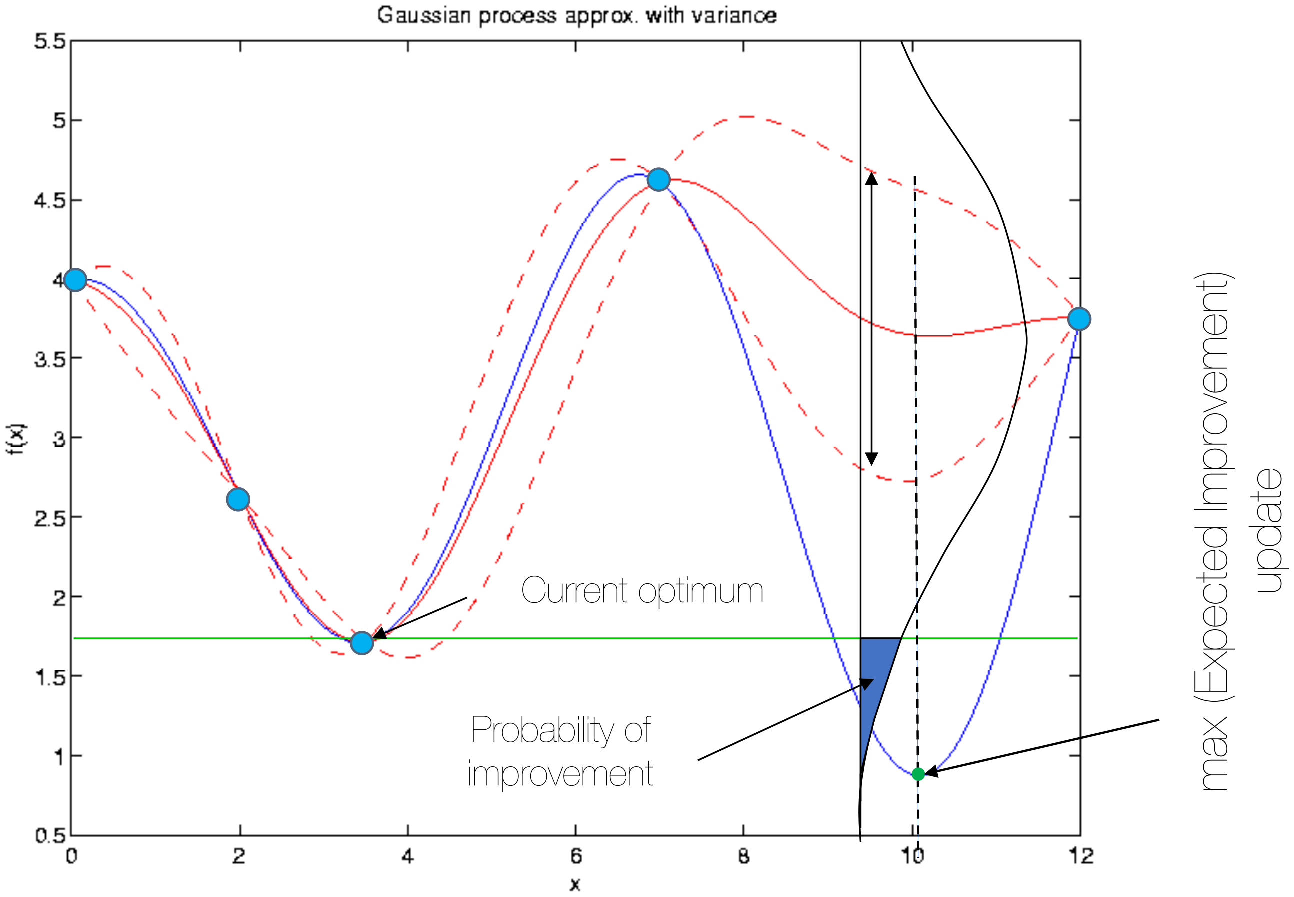


Illustration on 1D example

$$\begin{cases} \min \ (6x-2)^2 \sin(12x-4) \\ s.t. \\ 0 \le x \le 1 \end{cases}$$

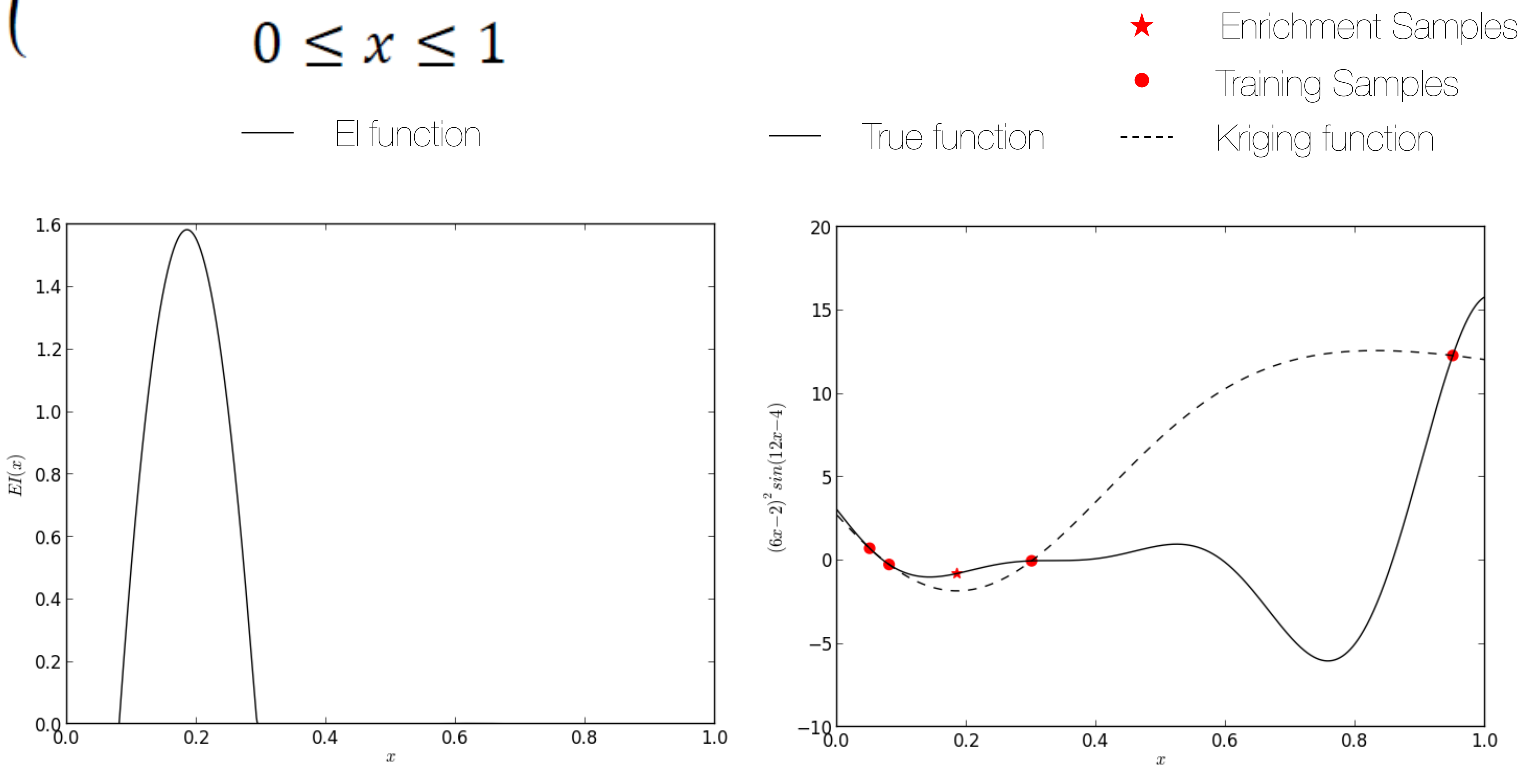


Illustration on 1D example

$$\begin{cases} \min \ (6x-2)^2 \sin(12x-4) \\ s.t. \\ 0 \le x \le 1 \end{cases}$$

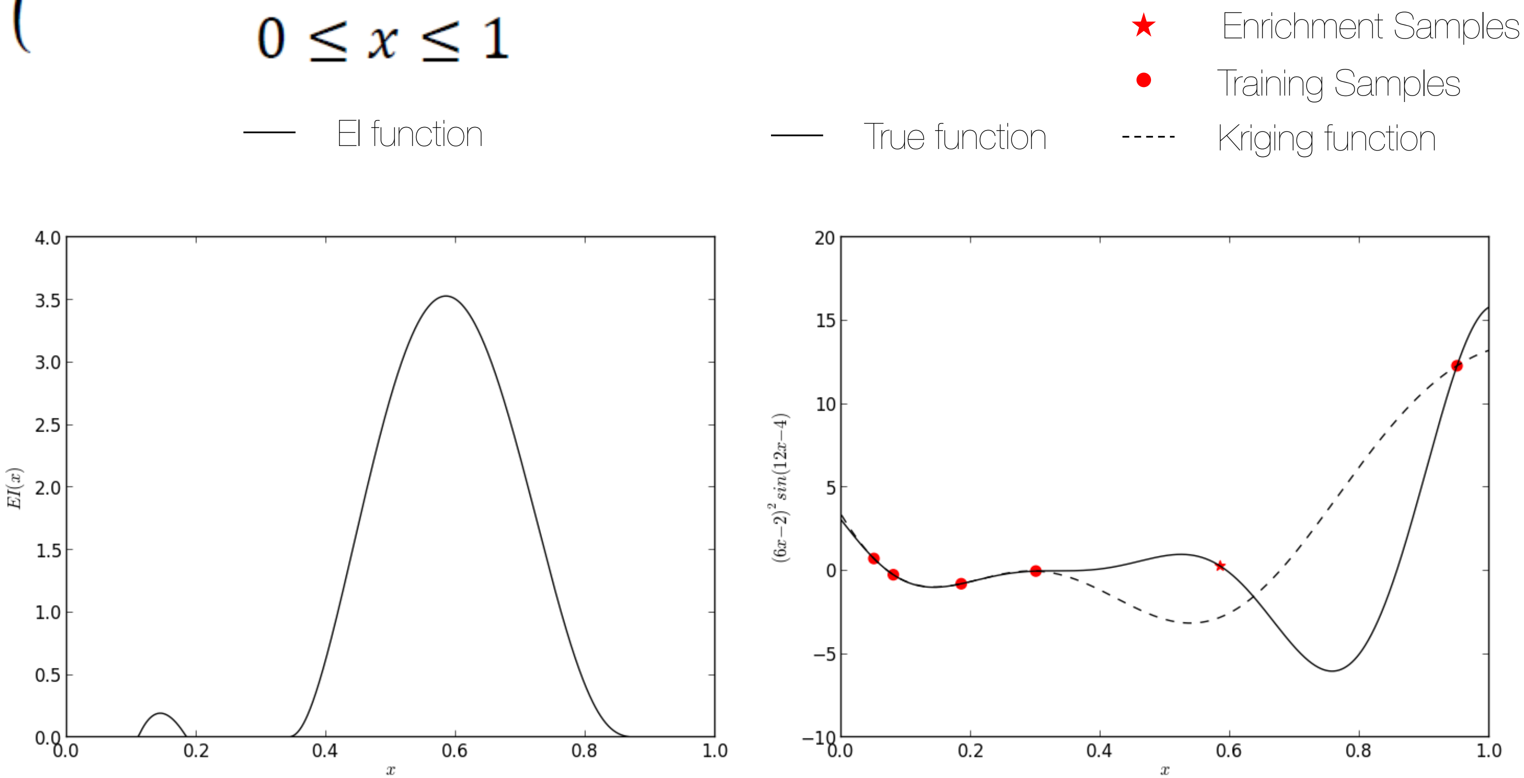


Illustration on 1D example

$$\begin{cases} \min \ (6x-2)^2 \sin(12x-4) \\ s.t. \\ 0 \le x \le 1 \end{cases}$$

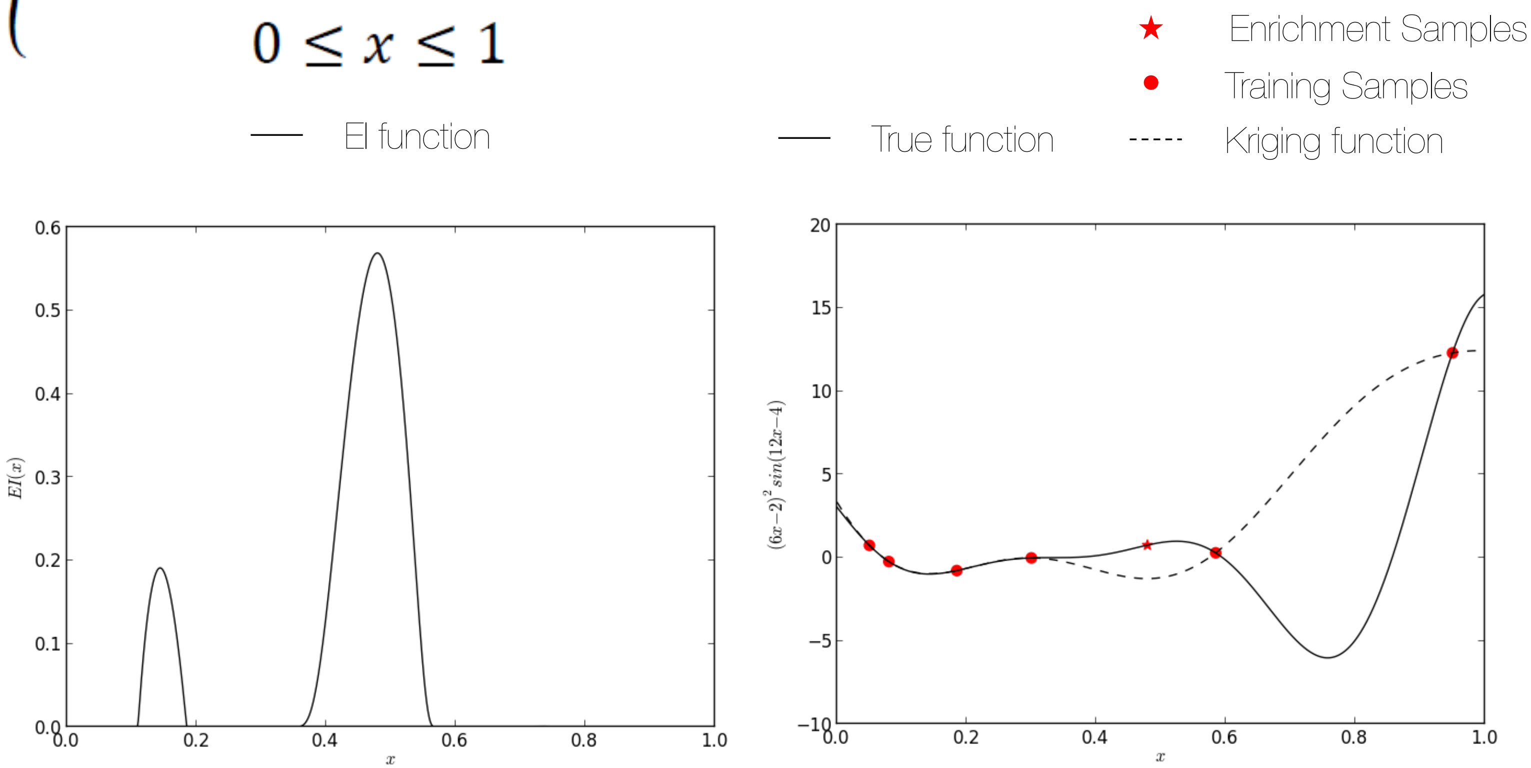
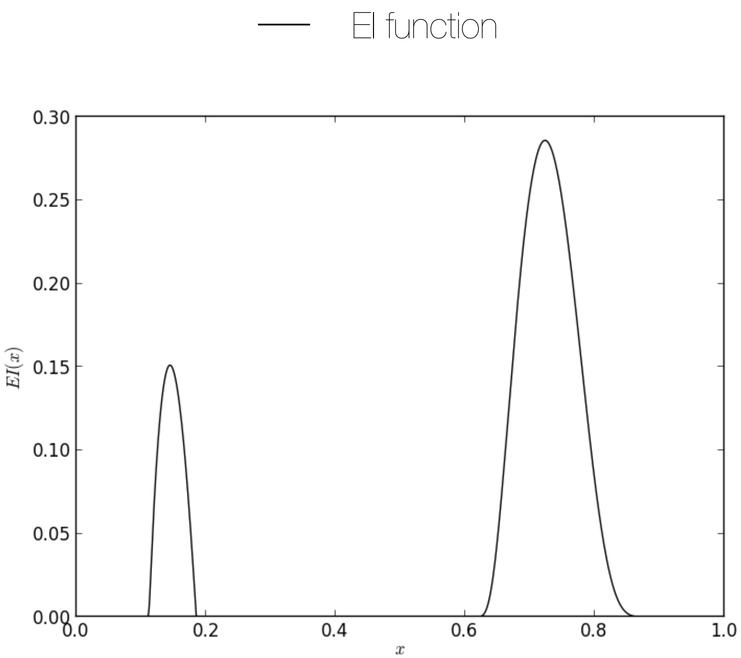


Illustration on 1D example

$$\begin{cases} \min \ (6x-2)^2 \sin(12x-4) \\ s.t. \\ 0 \le x \le 1 \end{cases}$$



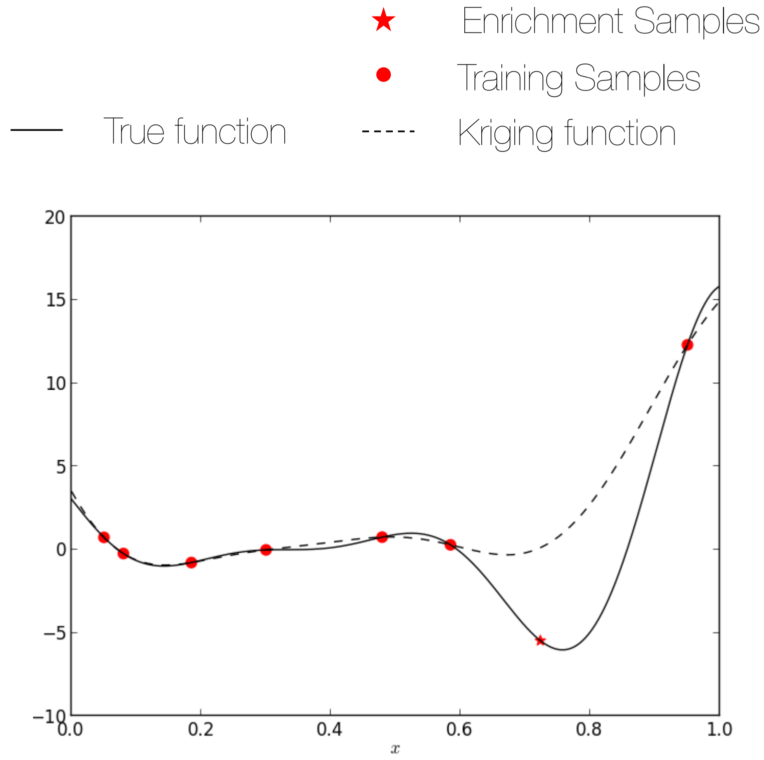
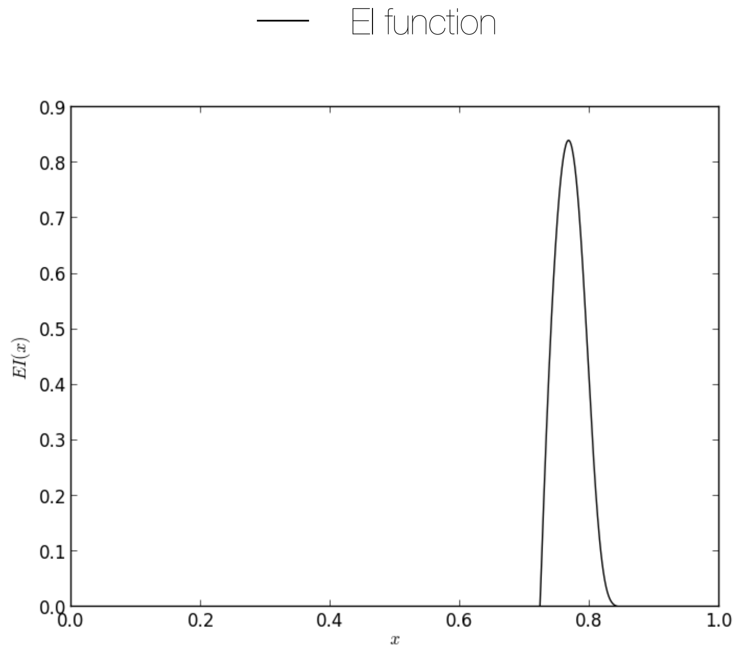
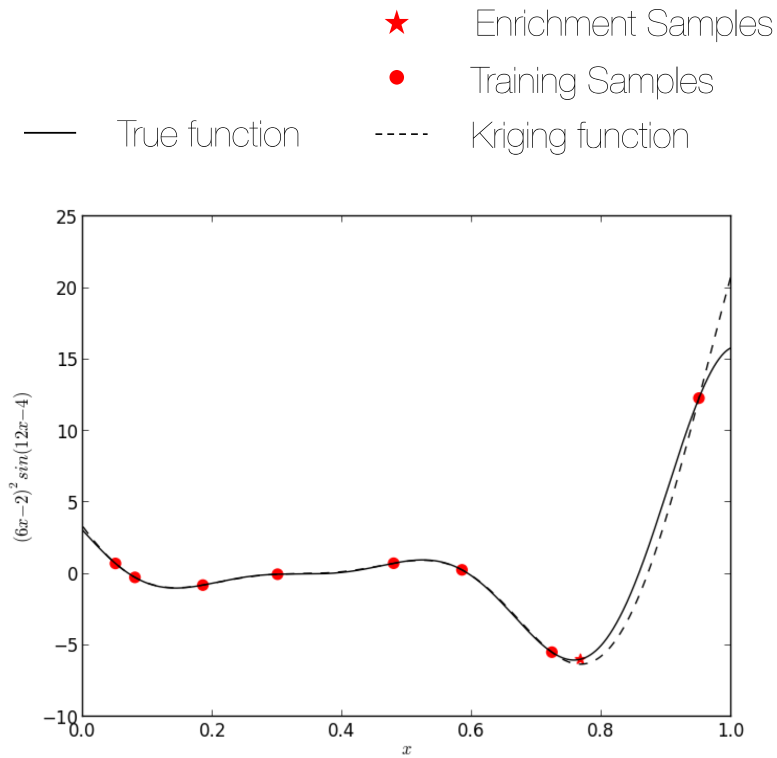


Illustration on 1D example

$$\begin{cases} \min \ (6x-2)^2 \sin(12x-4) \\ \qquad s.t. \\ \qquad 0 \le x \le 1 \end{cases}$$

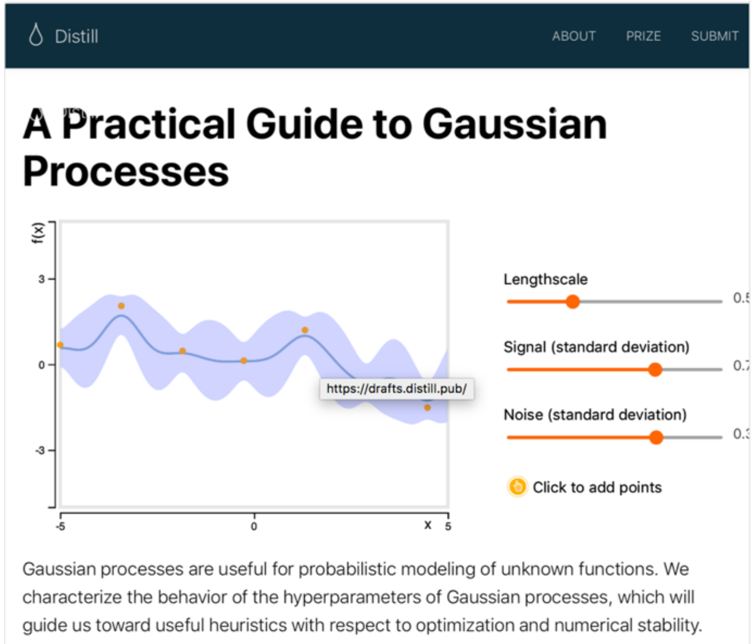




A good starting point x_0=Rasmussen’s book (ML)

A good starting point x_0=Forrester’s book (Aerospace)

- https://drafts.distill.pub/gp/



C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006, ISBN 026218253X. www.GaussianProcess.org/gpml

Gaussian Processes for Machine Learning

Engineering Design via Surrogate Modelling

A Practical Guide

Alexander I. J. Forrester, András Sóbester and Andy J. Keane
University of Southampton, UK